03/16/2020

RUpesh Chaturvedi

ETL Project

# Data Selection AND Extraction

Two Data sets are used from two different sources.

**UNICEF:** For immunization in each country as reported by respective administrative unit of the country from 1997–2018. Immunization coverage by antigen (country, regional, and global trends): BCG, DTP1, DTP3, HEPB3, HEPBB, HIB3, MCV1, MCV2, PCV3, POL3, RCV1, ROTAC

Data: <https://data.unicef.org/topic/child-health/immunization/>

**WHO:** For infection incidence preventable by vaccination data includs following infections for each country from 1997–2018. Diphtheria, Japanese encephalitis, Measles, Pertussis, Polio, Rubella, Tetanus, Yellow fever Data: (<http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tsincidencediphtheria.html>)

# Extraction

Data were in multiple sheets. Data were extracted using code to read from multiple worksheet.

# Transformation

Data were transformed only for 2011-2018 year. Columns related with other year were dropped by making new tables including only desired column headings. Rows with NaN values were dropped.

# Load

Further data were loaded into Relational data base using pgAdmin as Immunization\_infection\_db databases. Country\_code is primary key for both tables.